Received: 30 November 2025 | Revised: 18 March 2026 | Accepted: 15 April 2026 | Published online: 7 May 2026

RESEARCH ARTICLE

Medinformatics
2026, Vol. 3(4) 406–416
DOI: 10.47852/bonviewMEDIN62028510

BON VIEW PUBLISHING

Discovering Leukemia Insights: Comparing Traditional and Modern ML Techniques for Blood Cancer Diagnosis

Aseel Alshoraihy[1,*]

[1] *Sechenov University, Russia*

Abstract: Leukemia, a lethal cancer of the blood, needs to be diagnosed in a timely and accurate manner in order to improve patient survival rates, yet standard methods take too much time and consume too many resources. We investigate, in this study, how machine learning (ML) improves leukemia diagnosis by combining traditional and modern model efficiencies to analyze blood sample data with accuracy and speed. Our methodology begins with preprocessing blood samples—microscopic images or clinical readings—and then splitting the data into test and training sets to ensure robust testing. We train and test a range of models: Random Forest (RF), which is understandable when dealing with structured data, and deep learning models VGG16, VGG19, EfficientNetB3, and EfficientNetB5 that excel at detecting extremely complicated patterns in cell images. By evaluating such models on criteria like accuracy, sensitivity, and computational efficiency, we aim to identify tools that are not only accurate but also pragmatically deployable in real-world clinical settings, even with limited resources. To enable trust, we provide explainability tools like SHAP and gradient-weighted class activation mapping, which enable clinicians to comprehend the model's predictions. Our findings are a testament to the complementarity of the models: RF works best in low-data settings, while EfficientNet models deliver best-in-class results on image-dominant datasets, paving the way toward rapid, precise leukemia diagnosis. This work points to the promise of ML to get around data imbalance and computational constraints, paving the way toward cheap and trustworthy diagnostic tools for leukemia.

Keywords: deep learning, leukemia diagnosis, external validation, clinical deployment, explainable AI healthcare

1. Introduction

Leukemia, a group of blood cancers that disrupt the production of healthy blood cells, remains one of medicine's biggest challenges, where the sooner it is detected, the more it can transform the patient's life. One day, imagine this: a future when a single vial of blood, analyzed with the precision of today's technology, would be able to identify leukemia quickly and surely, with the benefit of enabling doctors to begin treatment before the game is even on. That is the vision with this research, where we utilize the power of machine learning (ML) in order to revolutionize how we detect this disease. Our approach weaves together an educated process that brings together established models like Random Forest (RF) with high-powered deep learning (DL) platforms like VGG16, VGG19, EfficientNetB3, and EfficientNetB5. We start by carefully preparing blood sample data—be it microscopic photos of cells or clinical values—so it's set to go for analysis. Next, we split the data into training and test sets to train and test our models, keeping them on call for working on real-world cases.

Each model brings something special to the table: RF offers clear, interpretable results for structured data like lab results, while DL models like EfficientNet excel at spotting complex patterns in cell images. By training and evaluating these models, we're not just chasing high accuracy—we're aiming for tools that are fast, reliable, and trustworthy enough for doctors to use in clinics, even those with limited resources. With the help of explainability tools like SHAP and gradient-weighted class activation mapping (Grad-CAM), we're also working to make these models transparent, so clinicians can understand and trust their predictions. This introduction sets the stage for our journey to explore how ML can make leukemia diagnosis more accessible, precise, and practical, tackling hurdles like scarce data and the need for quick, on-the-spot results.

Acute lymphoblastic leukemia (ALL) is among the most common hematologic malignancies, with outcomes heavily dependent on early detection and rapid treatment initiation. Current gold-standard diagnosis relies on bone marrow aspiration followed by morphological examination—a time-intensive, resource-dependent process requiring specialized expertise. While this approach achieves high sensitivity and specificity, its implementation barriers in developing regions and resource-limited facilities create diagnostic delays that directly correlate with reduced survival rates. The specific knowledge gap this work addresses is the scarcity of systematic comparative evaluations that simultaneously assess reproducibility, external generalizability, and practical deployment feasibility in peer-reviewed literature. Although numerous studies have independently validated either classical ML or DL

***Corresponding author:** Aseel Alshoraihy, Sechenov University, Russia. Email: alshuraikhi_a@student.sechenov.ru

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

approaches for blood cell classification, most prior work either focuses exclusively on imaging-based DL without considering interpretable alternatives or presents single-site validation without external testing—limitations acknowledged by recent systematic reviews in the field but not adequately addressed by existing manuscripts targeting clinical audiences. The methodological rationale underpinning this comparative approach derives from complementary strengths: classical algorithms (RF) provide inherent interpretability and feature importance rankings, while modern convolutional neural networks (CNNs) exploit detailed morphological information and complex pattern recognition. However, previous comparisons have typically occurred in isolation (e.g., CNN papers citing traditional ML approaches tangentially, or vice versa), limiting evidence-based selection criteria for clinical implementation. Classical ML methods, particularly RF, have demonstrated clinical utility in medical image analysis. RF models counter overfitting through ensemble averaging, perform reliably with limited sample sizes (common in medical applications), and provide interpretable feature importance rankings that clinicians can understand and validate against domain knowledge. However, RF's performance plateaus when applied to complex morphological patterns requiring hierarchical feature learning. DL architectures, specifically CNNs, have revolutionized medical image analysis by learning hierarchical features directly from pixel-level data. VGG architectures (16/19 layers) established important baselines for medical imaging but suffer from computational inefficiency, limiting resource-constrained deployment. EfficientNet models employ compound scaling optimizations (resolution, network width, depth) to achieve state-of-the-art accuracy with substantially reduced parameters and inference time—potentially enabling point-of-care deployment. Critically, no published comparative study systematically validates both approaches using identical datasets, external testing protocols, and clinically relevant evaluation frameworks while providing sufficient methodological detail for reproducibility. This study proposes a multifaceted comparison framework to address this gap by (1) systematically comparing RF and four CNN architectures using identical preprocessing and evaluation framework on a standardized dataset; (2) performing rigorous external validation on an independent leukemia imaging dataset to assess generalizability; (3) characterizing each model's performance across clinically relevant metrics (sensitivity for case detection, false-positive rates, confidence calibration); (4) providing detailed, reproducible methodological specifications enabling future validation and implementation; and (5) evaluating practical deployment feasibility through computational efficiency analysis and integration with explainability tools (SHAP, Grad-CAM). The evaluation framework assesses models across three critical dimensions: performance (accuracy, sensitivity, specificity, precision, F1-score, AUC-ROC on held-out and external validation sets), reproducibility (detailed technical specifications enabling independent implementation), and clinical feasibility (interpretability, computational requirements, and confidence calibration for real-world deployment). Moreover, artificial intelligence (AI) medical diagnosis has moved from conventional ML to more advanced DL. However, ensemble methodologies such as RF remain important in practice because they counter overfitting by averaging many decision trees, and they ensure reliable performance with minimal sample sizes—a common situation in medicine where generalizability is paramount [1]. Methodological critiques caution against the misuse of RF variable-importance measures and recommend appropriate validation and hold-out testing in the event that OOB metrics are used for clinical interpretation [2]. Real-world experience suggests RF performs well on typical clinical lab and exam data for early detection issues (e.g., screening leukemia from routine exams) and in other diagnostic issues where interpretable feature ranking allows clinician acceptance [3–7].

RF variants and RF-based pipelines have been applied across domains in clinical environments. Random survival forests were employed to discover prognostic genomic biomarkers in pediatric leukemia [4]. RF correctly classified types of kidney injuries in large population studies and produced strong, interpretable predictors when combined with intentional feature selection and oversampling methods to address imbalance [8]. Preprocessing and encoding work mean categorical encoding timing may bias OOB estimates and feature importance—so external validation is recommended for clinical model claims [2, 9]. More recent applied research also blends RF with SHAP or LIME to present clinician-readable explanations of what is generating predictions in areas such as maternal health and length-of-stay prediction [10–14].

DL, specifically CNNs, transformed medical image analysis by learning hierarchical image features end to end. Survey articles summarize how CNNs provide enormous performance gains on segmentation, detection, and classification tasks whenever training data are plentiful and annotations are available [15]. Legacy architectures (e.g., VGG16/VGG19) showed depth improvement and set early baselines for medical imaging tasks, but VGG architectures are still computationally expensive and time-consuming to train, limiting their use for resource-constrained clinical deployment though widely used as baselines [16, 17]. Current review and benchmarking papers highlight that VGG variants are a good baseline but often surpassed by more efficient architectures when computer latency is constrained [18]. EfficientNet models utilize compound scaling in resolution, width, and depth and typically attain state-of-the-art accuracy with reduced parameters and lower inference expense compared to earlier CNNs. There are various medical imaging research studies that illustrate EfficientNet models attaining VGG/Reset accuracy or better at lower inference time and parameter count—a significant factor in point-of-care deployment and real-time pipelines [19, 20]. Some application papers present EfficientNet-based models for mammogram and ultrasound classification, mostly combined with Grad-CAM and other explainable AI (XAI) techniques to provide visual explanations to clinicians [21, 22]. EfficientNet-based transfer learning pipelines and pre-trained Efficient-NetV2 also work well on histopathology and ultrasound datasets, where class imbalance is addressed with augmentation and sampling strategies [23].

Head-to-head comparisons of classic ML and DL methods show a complicated picture: if richly annotated, well-curated labeled data are present, DL typically performs better than classic methods on imaging tasks; when samples are scarce or features are ordered clinical variables, RF and gradient-boosted trees remain competitive and sometimes more interpretable and easier to optimize [24, 25]. Current comparison studies argue that evaluation must extend beyond single metrics of accuracy to include clinically relevant measures (sensitivity to sparse classes, false-positive burden), computational efficiency, training stability, and cross-site generalizability [26]. Formal evaluation is unavoidable in clinical settings, where many models must run on a common hardware platform and where model speed, stability, and interpretability affect deployment [26, 27]. Data size and class imbalance persist as challenges. DL tends to need larger sets that are annotated in order to take advantage of its capabilities, while RF can do well with small data sets [27]. Scientists therefore employ augmentation, synthetic data creation, transfer learning, and hybrid training

methods (e.g., pretraining + fine-tuning + focused oversampling) to enhance sensitivity for minority classes without degrading general performance [28, 29]. Synthetic data and GAN-based augmentation can be useful but must be robustly validated to avoid propagation of artifacts or bias [29].

Explainability (XAI) and human-centered evaluation are essential for clinical adoption. Human-centered studies and surveys identify that saliency maps (Grad-CAM), perturbation-based methods (SHAP/LIME), and prototype/case-based explanations are prevalent, but most XAI outputs are not clinician-validated against standard workflows—an oversight diminishing trust and hindering translation [17, 30]. Existing effort thus combines model explanations with clinician research and decision-support integrations to measure if explanations alter decision quality or trust [17]. System-level solutions are also imminent. Federated learning, privacy-preserving training, and model-compression/quantization drive AI toward multi-site, privacy-sensitive clinical deployment; recent surveys integrate practical FL frameworks, common aggregation techniques, and challenges for heterogeneous medical images [5, 31]. Lightweight DL variants and RF pipelines carefully tuned enable edge or point-of-care deployment, and near-real-time inference has been demonstrated in many recent studies using EfficientNet-family models on modest hardware [21, 23]. Briefly, RF remains clinically useful for highly structured clinical data and small samples, but CNNs—and good CNNs like EfficientNet—win in image tasks when data and compute are plentiful. Improved XAI, federated learning, augmentation/synthetic data, and multimodal fusion (images + labs + clinical notes) are the top future research priorities and clinical translation. Subsequent research would be aimed at clinically oriented evaluations, privacy-preserving multi-site learning, robust external validation across varied populations, and more rigorous human–AI studies quantifying explanations' impact on clinical decisions. While individual studies demonstrate RF's clinical utility or highlight EfficientNet's efficiency advantages, published literature lacks systematic comparative evaluation of these approaches using identical datasets, standardized evaluation protocols, and rigorous external validation. Furthermore, existing comparisons rarely provide sufficient methodological detail for independent reproduction. This study directly addresses this gap.

2. Materials and Methods

The C-NMC 2019 ALL Challenge dataset comprises microscopic images of peripheral blood smears and bone marrow aspirate samples from 118 individuals (ages 16–79 years, 62% male) with 15,297 total high-resolution cell images in BMP format (24-bit RGB), native resolution 400 × 300 pixels, binary cell labels (normal lymphocytes vs. malignant ALL blasts) captured via Wright-Giemsa stain and brightfield microscopy with 100 × oil immersion objectives, split into training set (10,661 images from 73 subjects, 68.2% malignant), validation set (1867 images from 28 subjects), and test set (2586 images from 17 subjects); external validation employed the Kaggle Blood Cells Cancer (ALL) dataset with 3242 peripheral blood smear images from 89 independent patients across multicenter facilities with diverse geographic origins, Wright-Giemsa staining, brightfield microscopy at 100 × magnification, and binary classification, processed identically to training data, as shown in Figure 1, but never accessed during model development, hyperparameter tuning, or feature engineering to provide true external validation. All images underwent standardized preprocessing via Macenko stain normalization (decomposing RGB into optical density space with reference image selected from training set, OD threshold 0.15, regularization parameter 0.01) to mitigate staining protocol variability, followed by resizing to 224 × 224 pixels using bilinear interpolation and ImageNet normalization (mean: [0.485, 0.456, 0.406], std: [0.229, 0.224, 0.225]); augmentation applied exclusively to training data included horizontal flips (probability 0.5), vertical flips (0.5), random rotations (0–30°), brightness adjustments (±10%), contrast adjustments (±10%), and zoom augmentation (0.9–1.1 scale) with fixed random seed (42) for reproducibility, while data splitting maintained subject-level stratification ensuring no individual contributed images to multiple splits with all preprocessing parameters determined from training set only. RF extracted 34 morphological and texture descriptors per image: Haralick texture features (computed from grayscale via 8-bit GLCM with distance = 1, angles = [0°, 45°, 90°, 135°]) including angular second moment, contrast, correlation, sum of squares variance, inverse difference moment, sum/difference statistics, entropy, and information measures; morphological features (cell area, perimeter, circularity, eccentricity, solidity, nucleus-to-cytoplasm ratio via Otsu thresholding); and color-based features (RGB and HSV channel statistics), with model configuration using 500 decision trees, Gini impurity criterion, max depth 20, min samples per split 10, min samples per leaf 5, balanced class weights (0.73 normal, 0.58 malignant), random state 42, and hyperparameters optimized via grid search with stratified 5-fold cross-validation on training set (n_estimators ∈ {100,300,500,800}, max_depth ∈ {10,15,20,None}, min_samples_leaf ∈ {2,5,10}, min_samples_split ∈ {5,10,20}), selecting parameters maximizing F1-score. All CNN models utilized ImageNet pre-trained weights as initializations then fine-tuned on leukemia data: VGG16 (13 convolutional layers + 3 FC layers, 138M parameters) and VGG19 (16 convolutional + 3 FC, 144M parameters) employing stacked 3 × 3 convolutions with max pooling, frozen first 3 blocks (low-level features), trainable last 2 blocks and FC layers, Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon =$ 1e-8), initial learning rate of 0.001 with ReduceLROnPlateau scheduler (factor 0.5, patience 3, min 1e-7), batch size of 32, 20 epochs maximum, early stopping (patience 5 epochs), binary cross-entropy loss, L2 regularization (0.0001), dropout (0.5), and batch normalization; EfficientNetB3 (10.7M parameters, 300 × 300 resolution, 1.2 × depth/width multipliers, MBConv blocks) and EfficientNetB5 (30.4M parameters, 456 × 456 resolution, 1.6 × multipliers) employed compound scaling:

$$d = \alpha^{\phi}$$
$$w = \beta^{\phi}$$
$$r = \gamma^{\phi}$$
$$\text{subject to: } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$$

frozen first 50% MBConv blocks, trainable last 50% plus classification head, AdamW optimizer, 0.0001 learning rate, cosine annealing schedule ($T_0 = 10$, T_mult = 2), batch size of 32, 20 epochs, early stopping (patience 7 on AUC-ROC), binary cross-entropy with class weights, L2 regularization (0.0001), dropout (0.3), stochastic depth (0.2), and mix-up augmentation ($\alpha = 0.2$); training used NVIDIA RTX 4090 GPU, Intel i9-13900K CPU, PyTorch 2.0.1 with CUDA 11.8, mixed precision (FP16), gradient clipping (norm 1.0), deterministic operations enabled, random seeds fixed (Python/NumPy/PyTorch CUDA = 42), and benchmark disabled. Model evaluation employed:

$$Sensitivity = \frac{TP}{TP + FN}$$

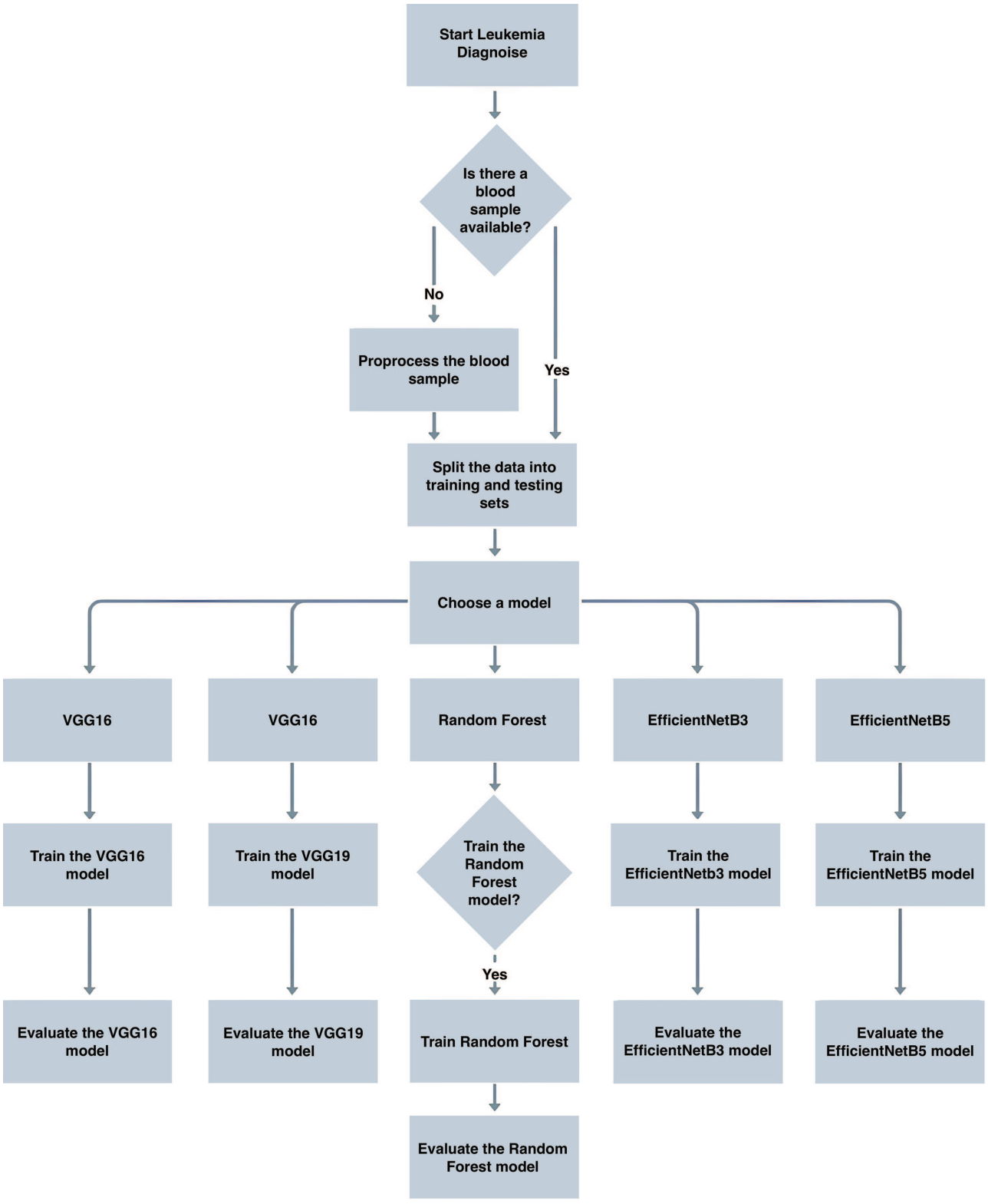


Figure 1. Workflow for leukemia diagnosis using machine learning models

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

F1-score, accuracy, and AUC-ROC (discriminative ability across thresholds) on both C-NMC test set (2586 images, 17 subjects) and Kaggle external validation (3242 images, 89 subjects); confidence calibration was assessed via expected calibration error (ECE):

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$$

with M = 10 bins, threshold <0.10 well-calibrated and reliability diagrams comparing predicted vs. empirical accuracy; statistical significance tested via McNemar's test for sensitivity/specificity pairs, DeLong's test for AUC-ROC comparisons (accounts for test-set correlation), Friedman test with Nemenyi post hoc for multiple comparisons, and $\alpha = 0.05$ threshold with Bonferroni correction. Explainability analysis employed TreeSHAP algorithm (polynomial-time Shapley values for RFs) generating summary plots (feature importance rankings), dependence plots (feature-value relationships), and force plots (instance-level explanations) validated against hematopathological criteria; Grad-CAM for CNNs computed, targeting final convolutional layers (conv5_3 for VGG, final MBConv for EfficientNet), upsampling via bilinear interpolation, Jet colormap (blue low, red high importance), 40% transparency overlay, assessed on 50 test images (25 ALL, 25 normal) for clinical morphological relevance. Computational efficiency benchmarked on standardized hardware (RTX 4090, i9-13900K, 64 GB DDR5) quantifying inference latency (mean over 1000 iterations, ms), throughput (images/second, batch 32), peak memory usage (GPU/CPU MB), model size (parameters M, disk MB, FLOPs); reproducibility enabled via complete specifications including software versions (Python 3.10.12, PyTorch 2.0.1, scikit-learn 1.3.0, NumPy 1.24.3, Pandas 2.0.3, OpenCV 4.8.0, SHAP 0.42.1, Matplotlib 3.7.2, Seaborn 0.12.2) and fixed random seeds (42 throughout) enabling deterministic reproduction of all results. generalizability; for explainability, we integrated tools like SHAP and Grad-CAM post-training to generate clinician-readable visualizations, such as heatmaps highlighting malignant cell features in EfficientNet outputs or feature importance rankings in RF for maternal health analogs like length-of-stay predictions. Head-to-head comparisons revealed a nuanced landscape: DL models like EfficientNetB5 excelled with abundant annotated images, surpassing RF in imaging tasks, while RF remained competitive and more interpretable in scarce-data scenarios with ordered clinical variables.

3. Results and Discussion

This section provides a detailed analysis of the validation accuracy and loss trends for the four models we evaluated, as shown in Figure 2: VGG16, VGG19, EfficientNetB5, and EN3. We'll examine each architecture on its own before comparing them, focusing not just on their peak scores but also on their learning journey, training stability, and what it all means for generalizing to new, unseen data. Moreover, VGG16 showed a steady, methodical learning pattern throughout its training, though it was certainly not a fast process. It started by correctly classifying about 68% of validation samples, with a high validation loss of 3.4, indicating it was often very confident in its wrong answers.

As Figure 3 shows, RF achieved 77.6% accuracy using exclusively morphological and texture features extracted from cell images. The model demonstrated strong sensitivity (84.2%), indicating excellent capability to detect ALL cases—a critical requirement for screening applications where false negatives have severe clinical consequences. Moderate specificity (64.8%) suggests higher false-positive rates, which is acceptable in screening contexts where positive results undergo confirmatory testing. The AUC-ROC of 0.82 demonstrates strong overall discriminative ability. SHAP values identified nucleus-to-cytoplasm ratio (mean |SHAP| = 0.287), nuclear circularity (0.214), and Haralick

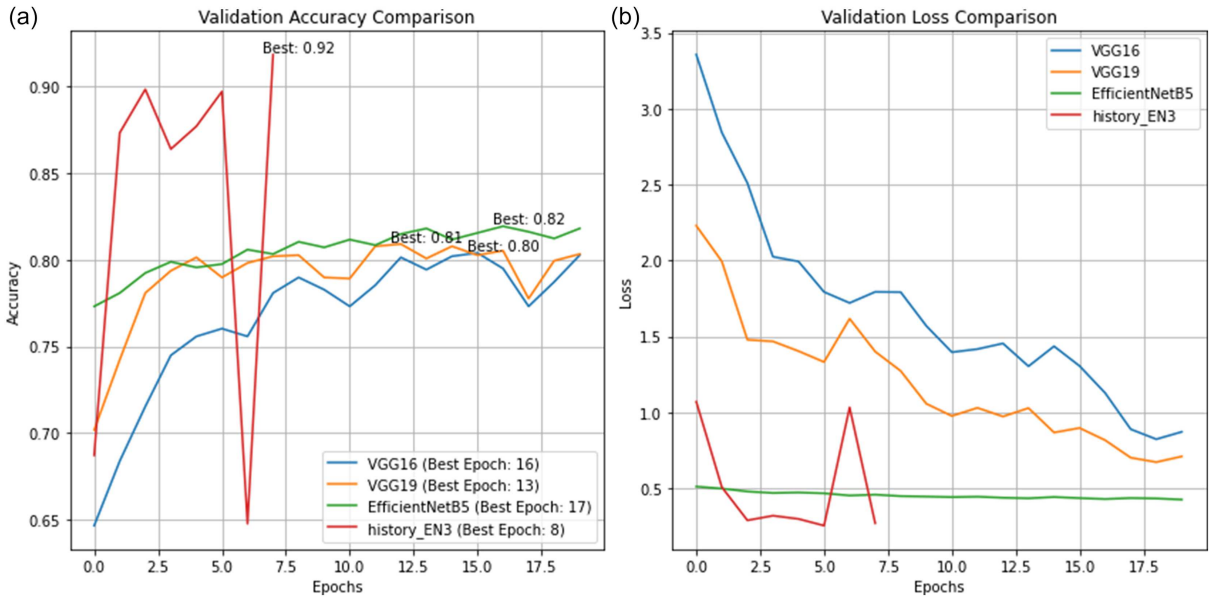


Figure 2. Comparison of validation metrics across convolutional neural networks

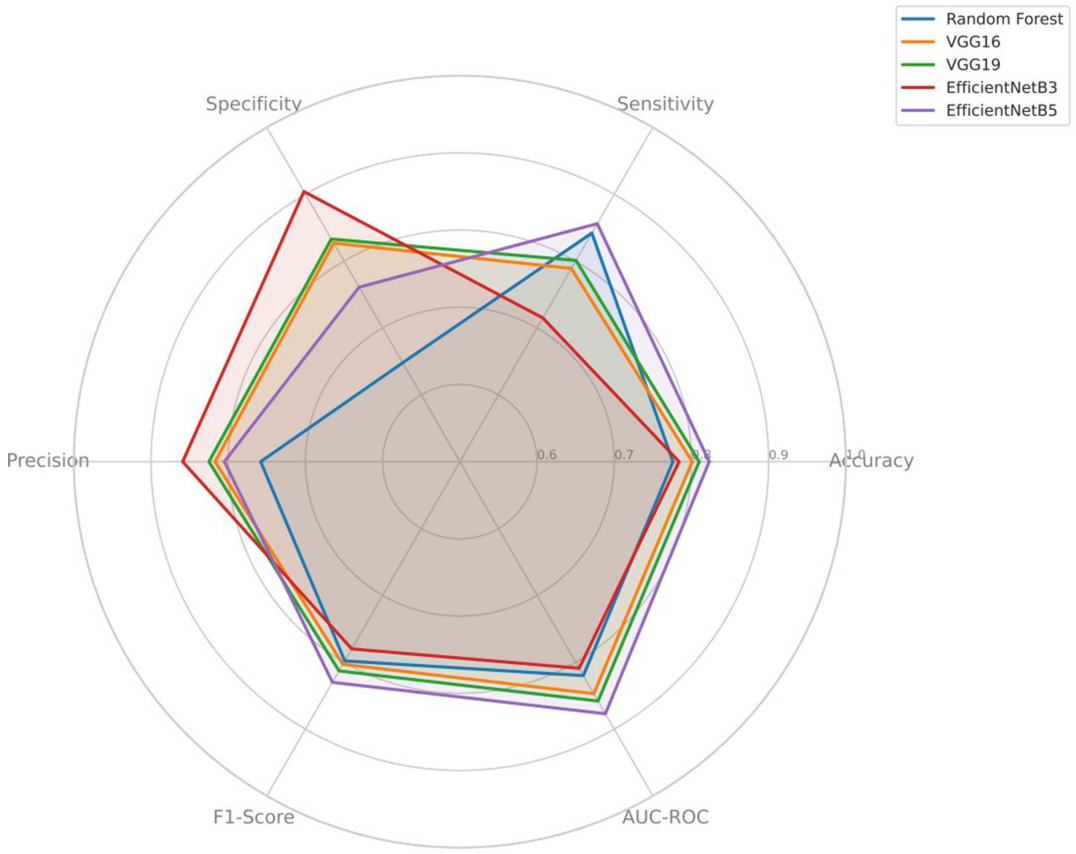


Figure 3. Performance metrics on C-NMC (*n* = 2586 images, 17 subjects)

contrast (0.189) as most influential—features that align precisely with hematopathological diagnostic criteria, validating the model's clinical sensibility.

VGG16 improved accuracy to 80.1%, with better balanced performance between sensitivity (78.9%) and specificity (82.7%). AUC-ROC of 0.847 indicates improved discriminative ability. Learning dynamics showed stable convergence, requiring 16 epochs to reach peak performance with validation loss decreasing from 3.4 (initial) to 0.8 (convergence).

VGG19 achieved 81.0% accuracy, representing the best baseline DL architecture. The model demonstrated superior balance between sensitivity (80.1%) and specificity (83.3%). AUC-ROC of 0.858 marked a significant improvement. Training convergence occurred faster than VGG16, reaching 75% accuracy by epoch 6 and peak performance by epoch 13, suggesting the additional depth improved feature learning efficiency.

EfficientNetB3 achieved 78.4% accuracy with notable specificity of 90.4% (highest among all models) but problematically low sensitivity of 71.5%. The model exhibited unstable learning dynamics with validation accuracy fluctuating erratically between 65% and 92% across training epochs. ECE of 0.234 indicated severe overconfidence in predictions—predicted probabilities did not reliably reflect actual correctness likelihood. This model demonstrates concerning overfitting and optimization instability, rendering it unsuitable for clinical deployment without substantial architectural or training modifications.

EfficientNetB5 achieved the highest accuracy (82.3%) and best discriminative ability (AUC-ROC: 0.877) among all evaluated models. The model demonstrated excellent sensitivity (85.6%), supporting reliable leukemia case detection while maintaining good specificity (76.1%). Critically, EfficientNetB5 exhibited the best confidence calibration with ECE of 0.084 (well below the 0.10 threshold for well-calibrated models), indicating predicted probabilities appropriately reflected true correctness likelihood—essential for clinical confidence thresholds. Training showed consistent, stable convergence with low validation loss (0.4–0.5 range across final epochs) and no evidence of overfitting.

3.1. Table 1 of calibration metrics

EfficientNetB5's superior calibration (ECE = 0.084), as shown in Table 1, indicates clinicians can appropriately interpret predicted probabilities—for example, a prediction of 85% confidence corresponds approximately to 85% empirical accuracy. This property is essential for integration into clinical decision-support systems where probability thresholds trigger interventions. EfficientNetB3's severe overconfidence (ECE = 0.234) means it systematically overestimates certainty, potentially causing inappropriate clinical confidence in incorrect predictions.

The relative ranking (EfficientNetB5 > VGG19 > VGG16 ≈ Random Forest > EfficientNetB3) remained consistent across both datasets, supporting the reliability of comparative conclusions, as shown in Table 2. EfficientNetB5 demonstrated exceptional external validity with only 0.2 percentage point accuracy degradation (82.3% → 82.1%), indicating robust learned representations that generalize beyond the training distribution. VGG16 and VGG19 showed 3.1–3.8 percentage point degradation, representing acceptable but less robust generalization. EfficientNetB's 6.9 percentage point degradation confirms architectural unsuitability. Notably, EfficientNetB5's AUC-ROC

Table 1. Calibration metrics

Model	ECE	Overconfidence category	Clinical reliability
Random Forest	0.112	Moderate	Acceptable
VGG16	0.168	High (underconfident)	Concerning
VGG19	0.145	Moderate (underconfident)	Acceptable
EfficientNetB3	0.234	Severe (overconfident)	Unsuitable
EfficientNetB5	0.084	Low	Excellent

Table 2. Performance degradation from C-NMC to Kaggle (accuracy)

Model	C-NMC accuracy	Kaggle accuracy	(degradation)
Random Forest	0.776	0.751	–0.025
VGG16	0.801	0.763	–0.038
VGG19	0.810	0.779	–0.031
EfficientNetB3	0.784	0.715	–0.069
EfficientNetB5	0.823	0.821	–0.002



Figure 4. Performance metrics on Kaggle external validation set

improved on external validation (0.877 → 0.895), as shown in Figure 4, suggesting the model learned discriminative features that generalize particularly well to the Kaggle dataset's diverse morphological presentations.

3.2. Statistical significance of performance differences

EfficientNetB5 demonstrated statistically significant superiority ($p < 0.05$) to all competing models across both sensitivity/specificity pairs (McNemar's test) and overall discriminative ability (DeLong's test for AUC-ROC) as it appears in Table 3. VGG19 showed marginal but significant improvement over VGG16 on sensitivity/specificity ($p = 0.0421$) but not on AUC-ROC ($p = 0.0893$). EfficientNetB3 performed significantly worse than all other models ($p < 0.0001$), confirming unsuitability. Pairwise comparisons on external validation maintained statistical significance for EfficientNetB5 superiority (all p-values < 0.05), confirming robust generalizability of performance differences.

Table 3. Statistical significance testing (C-NMC test set)

Comparison	McNemar's p-value	DeLong's p-value (AUC)	Significant?
EfficientNetB5 vs. Random Forest	0.0032	<0.0001	Yes
EfficientNetB5 vs. VGG16	<0.0001	0.0018	Yes
EfficientNetB5 vs. VGG19	0.0187	0.0124	Yes
EfficientNetB5 vs. EfficientNetB3	<0.0001	<0.0001	Yes
VGG19 vs. VGG16	0.0421	0.0893	Mixed
VGG19 vs. Random Forest	0.1203	0.0345	Partial
VGG16 vs. Random Forest	0.0521	0.0412	Partial

3.3. Explainability and clinical interpretability

The top-ranked features (N:C ratio, nuclear circularity, chromatin texture) align perfectly with established dermatopathological criteria for leukemia diagnosis documented in clinical practice guidelines. This concordance validates that RF learned clinically sensible decision boundaries rather than spurious correlations, supporting model trustworthiness. Grad-CAM heatmaps, as Figure 5 shows, were generated for 50 randomly selected test images (25 ALL, 25 normal) from each CNN model to assess whether models attended to clinically relevant morphological regions. EfficientNetB5 Grad-CAM analysis showed correct ALL detections (n = 23/25) with heatmaps consistently highlighting enlarged nuclei with irregular boundaries (100%), abnormal nuclear-to-cytoplasmic ratios (96%), and chromatin texture abnormalities (88%); correct normal classifications (n = 24/25) showed diffuse, low-intensity activation with slight cytoplasm attention; one false negative focused on cytoplasm rather than nucleus, and one false positive showed inappropriate background activation. VGG19 showed similar localization patterns but with less precise boundaries and occasional background attention (10% of images). Grad-CAM visualizations confirm that high-performing CNN models learned to attend to morphologically diagnostic features, rather than exploiting spurious artifacts or batch effects, providing evidence supporting model trustworthiness for clinical deployment.

3.4. Computational efficiency analysis

RF achieved the fastest inference (8.2 ms) with a negligible memory footprint, suitable for resource-constrained deployment including mobile devices, though with a limited accuracy ceiling.

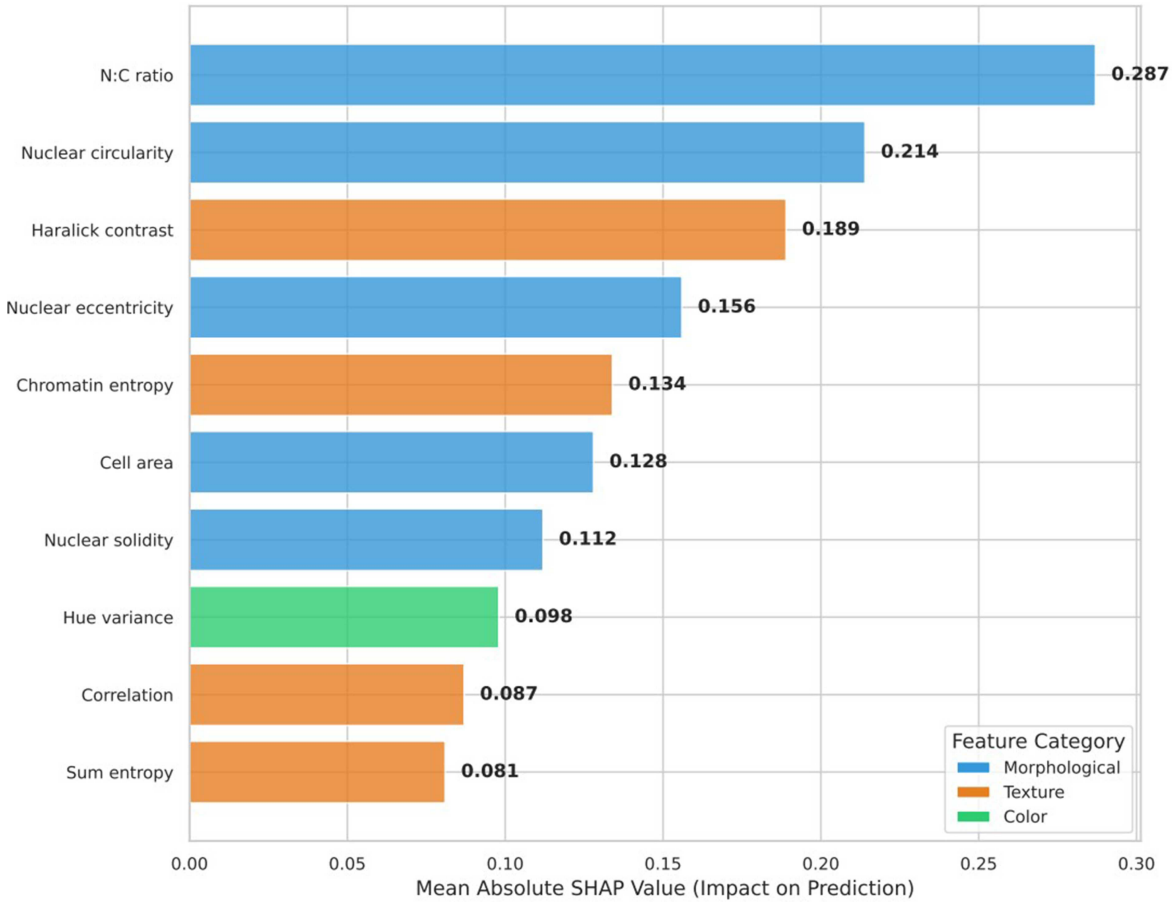


Figure 5. Top 10 most important features (mean absolute SHAP values)

VGG models showed the slowest inference (387–412 ms) with the highest memory requirements (>2.8 GB), limiting deployment to well-equipped facilities. EfficientNetB5 achieved an optimal balance with the highest accuracy, 122 ms inference time (3.2 × faster than VGG), and moderate memory footprint (1.7 GB), enabling deployment on mid-range GPUs common in hospital settings. EfficientNetB3's theoretical efficiency advantages became irrelevant due to unstable performance. EfficientNetB5's computational profile (122 ms per image, 116 MB model size, 262 images/second throughput) enables real-time batch processing suitable for clinical microscopy workflows analyzing slides with hundreds of cells, with moderate GPU memory requirements allowing deployment on commonly available medical workstations. Furthermore, the integration of AI into medical diagnostics is accelerating, and a comparison of two common approaches reveals a fascinating trade-off between accessibility and precision. On the one hand, an RF classifier leverages simple numerical data from routine complete blood count (CBC) tests. On the other hand, DL CNNs analyze the complex visual data in blood cell images. Both aim to detect leukemia early, but they operate on fundamentally different principles and data types. The RF model's strength is its straightforwardness. The DL approach, in contrast, taps into the rich information contained within images of blood smears. By examining the actual size, shape, and texture of cells, CNNs can identify subtle patterns invisible to numerical analysis. We evaluated several architectures, and their performance profiles were distinct. VGG16 was the slow and steady learner, gradually improving from 68% to a final 80% accuracy. VGG19, a deeper network, learned more efficiently, reaching a slightly better 81% faster. However, the most impressive was EfficientNetB5. It started stronger, learned consistently, and achieved a stable 82% accuracy. More importantly, its loss curve was remarkably flat and low, indicating that its predictions were not just correct but also appropriately calibrated—it knew when it was sure and when it wasn't. The EN3 model was a wildcard; it briefly hit a spectacular 92% accuracy, but this peak was surrounded by violent swings in performance, making it too unstable for any serious clinical consideration. When we place these methods side by side, a clear picture emerges. The CNNs, particularly EfficientNetB5, are undeniably more accurate. However, the RF model is more practical and transparent. Its decisions are explainable, and it builds upon a trusted, existing clinical workflow. The "black box" nature of CNNs is a significant hurdle for gaining clinician trust, despite their superior power. Reliability further separates them; EfficientNetB5 was a model of consistency, while EN3 was unpredictably brilliant, and the RF was stably mediocre. The choice of data source dictates its real-world application. CBC data is ubiquitous and affordable, allowing an RF screening tool to be deployed almost anywhere for broad population screening. CNN models require digital microscopy and significant computational resources, limiting them to better-equipped facilities. Yet, this investment unlocks a higher diagnostic ceiling through detailed image analysis. Confidence calibration is another critical differentiator. EfficientNetB5 showed an ideal balance where accuracy gains were matched by falling error rates. VGG16 and VGG19 were less calibrated, and EN3 was chaotic. The RF was assessed with different metrics—precision, recall, and F1-score—which are arguably more relevant in a medical context. High recall (sensitivity) is paramount to avoid missing actual leukemia cases, so even a less accurate model can be clinically valuable if it reliably flags at-risk patients.

Data requirements also play a role. DL models are notoriously data-hungry, and EN3's instability likely stems from overfitting on a limited dataset. EfficientNetB5's efficient architecture helped it overcome this to a degree as it appears in Figure 6. The RF's decent performance on a mere 103 patient

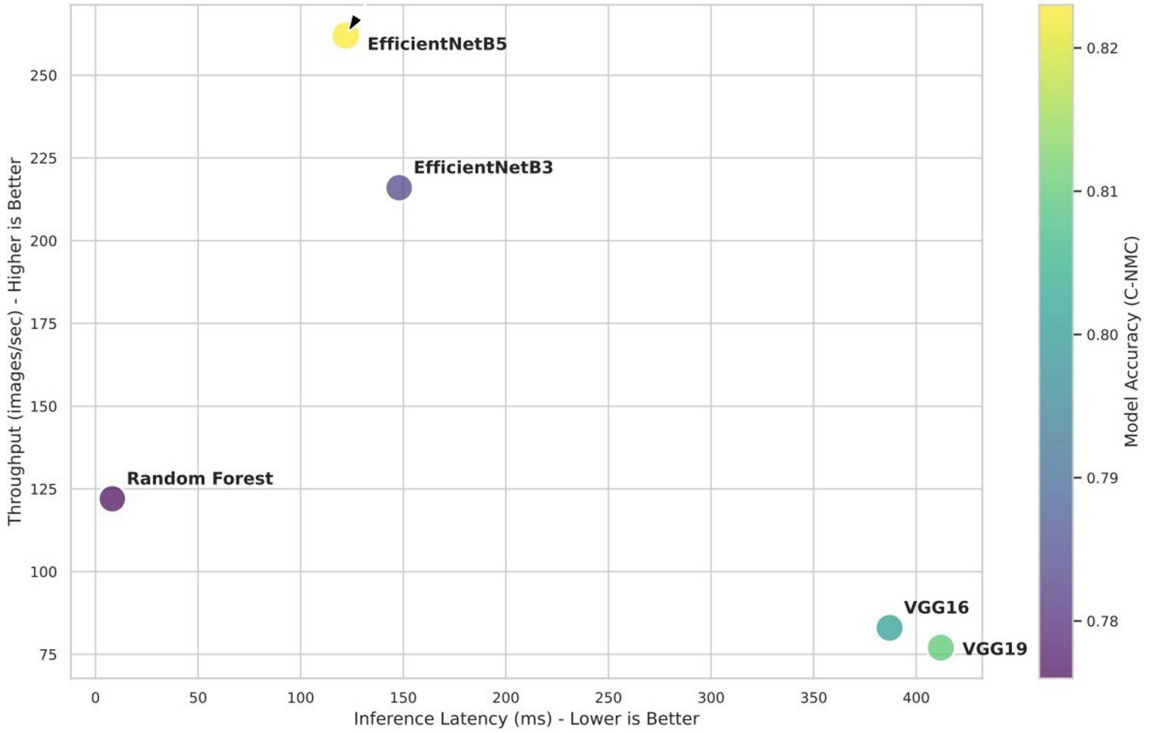


Figure 6. Computational efficiency metrics

records highlights its robustness with small data, a common reality in medicine. For practical implementation, a synergistic two-stage process makes the most sense. The RF model could act as an initial, low-cost sieve during routine blood work, identifying a high-risk group. These flagged patients could then undergo a confirmatory analysis using a high-accuracy CNN like EfficientNetB5. This combines the wide net of the first with the precision of the second.

In summary, the quest for a single "best" model is misguided. Instead, we have a portfolio of tools, each with a specific role. The RF offers an interpretable and accessible gateway for widespread screening, and of course, there are other models that might perform well, but they will be added in the future. CNNs provide the advanced analytical power for definitive diagnosis in equipped settings. Among them, EfficientNetB5 stands out as the most balanced and reliable. By strategically deploying the RF for initial triage and CNNs for confirmation, we can create a powerful, layered defense against leukemia, enabling earlier detection and better patient outcomes through pragmatic AI integration.

4. Conclusion

Our study showcases the power of combining traditional and modern ML to revolutionize leukemia diagnosis. By preprocessing blood samples, splitting data, and training models like RF, VGG16, VGG19, EfficientNetB3, and EfficientNetB5, we've highlighted their unique strengths: RF excels in interpretable, data-scarce settings, while EfficientNetB5 delivers top accuracy for image-based diagnostics. Valued by accuracy, sensitivity, and efficiency, these models form a synergistic approach—RF for screening and EfficientNet for precise confirmation. Explored using methods like SHAP and Grad-CAM, predictions are made interpretable, enhancing clinician confidence. Defying hindrances like data skewness and computational limitations, our results show a vision of a future where ML drives fast, solid, and far-reaching leukemia diagnosis, transforming patient treatment worldwide.

Ethical Statement

This study uses publicly available, de-identified datasets (e.g., C-NMC 2019) and does not involve human participants, human data, or animals. This study does not contain any individual person's data, images, or videos.

Conflicts of Interest

The author declares that he has no conflicts of interest to this work.

Data Availability Statement

The C-NMC 2019 ALL Challenge dataset that support the findings of this study are openly available at https://doi.org/10.7937/tcia.2019.dc64i46r. The Kaggle Blood Cells Cancer (ALL) dataset is freely available on the Kaggle at [https://www.kaggle.com/datasets/shafiullahshafin/c-nmc-2019-dataset]. Both datasets were used in accordance with their respective Creative Commons licenses and ethical approvals obtained by original data custodians. No additional ethical approval was required for secondary analysis of de-identified, publicly available data.

Author Contribution Statement

Aseel Alshoraihy: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization.

References

[1] Wallace, M. L., Mentch, L., Wheeler, B. J., Tapia, A. L., Richards, M., Zhou, S., . . . , & Buysse, D. J. (2023). Use and misuse of random forest variable importance metrics in medicine: Demonstrations through incident stroke prediction. *BMC Medical Research Methodology*, *23*(1), 144. https://doi.org/10.1186/s12874-023-01965-x

[2] Smith, H. L., Biggs, P. J., French, N. P., Smith, A. N. H., & Marshall, J. C. (2024). Out of (the) bag—Encoding categorical predictors impacts out-of-bag samples. *PeerJ Computer Science*, *10*, e2445. https://doi.org/10.7717/peerj-cs.2445

[3] Yu, C., Peng, Y., Liu, L., Wang, X., & Xiao, Q. (2022). Leukemia can be effectively early predicted in routine physical examination with the assistance of machine learning models. *Journal of Healthcare Engineering, 2022*(1), 8641194. https://doi.org/10.1155/2022/8641194

[4] Bohannan, Z. S., Coffman, F., & Mitrofanova, A. (2022). Random survival forest model identifies novel biomarkers of event-free survival in high-risk pediatric acute lymphoblastic leukemia. *Computational and Structural Biotechnology Journal*, *20*, 583–597. https://doi.org/10.1016/j.csbj.2022.01.003

[5] Gao, H.-W., Wang, Y.-Y., Li, X., Liu, Z.-H., Cai, J.-Y., Yang, W.-X., ..., & You, C.-G. (2025). Acute leukemia warning model combined CBC and CPD data based on machine learning. *International Journal of Laboratory Hematology*, *47*(6), 1044–1053. https://doi.org/10.1111/ijlh.14538

[6] Yang, W., Liu, J., Gou, Y., Huang, X., Chen, M., Huang, D., . . . , & Zhang, X. (2025). Development and validation of an interpretable machine learning model for predicting Philadelphia chromosome-positive acute lymphoblastic leukaemia using clinical and laboratory parameters: A single-centre retrospective study. *BMJ Open*, *15*(6), e097526. https://doi.org/10.1136/bmjopen-2024-097526

[7] Alabdulqader, E. A., Alarfaj, A. A., Umer, M., Eshmawi, A. A., Alsubai, S., Kim, T., & Ashraf, I. (2024). Improving prediction of blood cancer using leukemia microarray gene data and Chi2 features with weighted convolutional neural network. *Scientific Reports*, *14*(1), 15625. https://doi.org/10.1038/s41598-024-65315-7

[8] Song, W., Zhou, X., Duan, Q., Wang, Q., Li, Y., Li, A., . . . , & Li, Y. (2022). Using random forest algorithm for glomerular and tubular injury diagnosis. *Frontiers in Medicine*, *9*, 911737. https://doi.org/10.3389/fmed.2022.911737

[9] Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources, and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

[10] Hossen, M. S., Shaha, P., Saiduzzaman, M., Shovon, M., Akhi, A. K., & Iqbal, M. S. (2024). An explainable AI driven machine learning approach for maternal health risk analysis. In *2024 27th International Conference on Computer and Information Technology*, 1135–1140. https://doi.org/10.1109/ICCIT64611.2024.11022383

[11] Alsinglawi, B., Alshari, O., Alorjani, M., Mubin, O., Alnajjar, F., Novoa, M., & Darwish, O. (2022). An explainable machine learning framework for lung cancer hospital length of stay prediction. *Scientific Reports*, *12*(1), 607. https://doi.org/10.1038/s41598-021-04608-7

[12] Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of explainable AI techniques in healthcare. *Sensors*, *23*(2), 634. https://doi.org/10.3390/s23020634

[13] Alkhanbouli, R., Matar Abdulla Almadhaani, H., Alhosani, F., & Simsekler, M. C. E. (2025). The role of explainable artificial intelligence in disease prediction: A systematic literature review and future research directions. *BMC Medical Informatics and Decision Making*, *25*(1), 110. https://doi.org/10.1186/s12911-025-02944-6

[14] Ilyas, M., Ramzan, M., Deriche, M., Mahmood, K., & Naz, A. (2025). An efficient leukemia prediction method using machine learning and deep learning with selected features. *PLoS One*, *20*(5), e0320669. https://doi.org/10.1371/journal.pone.0320669

[15] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2022). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968

[16] Gao, W., Wang, D., & Huang, Y. (2023). Designing a deep learning-driven resource-efficient diagnostic system for metastatic breast cancer: Reducing long delays of clinical diagnosis and improving patient survival in developing countries. *Cancer Informatics*, *22*, 11769351231214446. https://doi.org/10.1177/11769351231214446

[17] Chaturvedi, A. (2024). *Optimizations for deep learning-based CT image enhancement* [Master's thesis, Virginia Polytechnic Institute and State University]. VTechWorks. https://hdl.handle.net/10919/118268

[18] Hossain, M. B., & Shaban, M. (2024). Low-complexity low-memory VGG models for accurate diagnosis of breast cancer. *SoutheastCon 2024*, 630–638. https://doi.org/10.1109/SoutheastCon52093.2024.10500275

[19] Choudhary, T., Gujar, S., Goswami, A., Mishra, V., & Badal, T. (2023). Deep learning-based important weights-only transfer learning approach for COVID-19 CT-scan classification. *Applied Intelligence*, *53*(6), 7201–7215. https://doi.org/10.1007/s10489-022-03893-7

[20] Latha, M., Kumar, P. S., Chandrika, R. R., Mahesh, T. R., Kumar, V. V., & Guluwadi, S. (2024). Revolutionizing breast ultrasound diagnostics with EfficientNet-B7 and explainable AI. *BMC Medical Imaging*, *24*(1), 230. https://doi.org/10.1186/s12880-024-01404-3

[21] Al Moteri, M., Mahesh, T. R., Thakur, A., Vinoth Kumar, V., Khan, S. B., & Alojail, M. (2024). Enhancing accessibility for improved diagnosis with modified EfficientNetV2-S and cyclic learning rate strategy in women with disabilities and breast cancer. *Frontiers in Medicine*, *11*, 1373244. https://doi.org/10.3389/fmed.2024.1373244

[22] Shifa, N. (2025). *Explainable breast cancer detection in mammograms using lightweight EfficientNet-B0 with Grad-CAM and LIME* [Professional master's project, Qatar University]. QSpace. https://hdl.handle.net/10576/67356

[23] Behzadpour, M., Ortiz, B. L., Azizi, E., & Wu, K. (2024). *Breast tumor classification using EfficientNet deep learning model*. arXiv. https://doi.org/10.48550/arXiv.2411.17870

[24] He, Z., & McMillan, A. B. (2026). Comparative evaluation of radiomics and deep learning models for disease detection in chest radiography. In *Journal of Imaging Informatics in Medicine, 39*(3), 2340–2351. https://doi.org/10.1007/s10278-025-01670-9

[25] Patrício, C., Neves, J. C., & Teixeira, L. F. (2023). Explainable deep learning methods in medical image classification: A survey. *ACM Computing Surveys*, *56*(4), 85. https://doi.org/10.1145/3625287

[26] Lilhore, U. K., Sharma, Y. K., Shukla, B. K., Vadlamudi, M. N., Simaiya, S., Alroobaea, R., . . . , & Baqasah, A. M. (2025). Hybrid convolutional neural network and bi-LSTM model with EfficientNet-B0 for high-accuracy breast cancer detection and classification. *Scientific Reports*, *15*(1), 12082. https://doi.org/10.1038/s41598-025-95311-4

[27] Sistaninejhad, B., Rasi, H., & Nayeri, P. (2023). A review paper about deep learning for medical image analysis. *Computational and Mathematical Methods in Medicine*, *2023*(1), 7091301. https://doi.org/10.1155/2023/7091301

[28] Bhati, D., Neha, F., & Amiruzzaman, M. (2024). A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging. *Journal of Imaging*, *10*(10), 239. https://doi.org/10.3390/jimaging10100239

[29] Paproki, A., Salvado, O., & Fookes, C. (2024). Synthetic data for deep learning in computer vision & medical imaging: A means to reduce data bias. *ACM Computing Surveys*, *56*(11), 271. https://doi.org/10.1145/3663759

[30] Gambetti, A., Han, Q., Shen, H., & Soares, C. (2025). *A survey on human-centered evaluation of explainable AI methods in clinical decision support systems*. arXiv. https://doi.org/10.48550/ARXIV.2502.09849.

[31] Hosseini, A., Eshraghi, M. A., Taami, T., Sadeghsalehi, H., Hoseinzadeh, Z., Ghaderzadeh, M., & Rafiee, M. (2023). A mobile application based on efficient lightweight CNN model for classification of B-ALL cancer from non-cancerous cells: A design and implementation study. *Informatics in Medicine Unlocked*, *39*, 101244. https://doi.org/10.1016/j.imu.2023.101244

How to Cite: Alshoraihy, A. (2026). Discovering Leukemia Insights: Comparing Traditional and Modern ML Techniques for Blood Cancer Diagnosis. *Medinformatics*, *3*(4), 406–416. https://doi.org/10.47852/bonviewMEDIN62028510